DOCUMENT RESUME

ED 185 076 TM 800 066

AUTHOR Kane, Michael T.; Brennan, Robert L.
TITLE Agreement Coefficients as Indices of Dependability for Domain-Referenced Tests. ACT Technical Bulletin No. 28.
INSTITUTION American Coll. Testing Program, Iowa City, Iowa. Research and Development Div.
SPONS AGENCY Navy Personnel Research and Development Center, San Diego, Calif.
PUB DATE Dec 77
CONTRACT N00123-77-C-0739
NOTE 58p.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Bayesian Statistics; Correlation; *Criterion Referenced Tests; *Cutting Scores; Guessing (Tests); *Mastery Tests; Mathematical Models; Norm Referenced Tests; Reliability; Statistical Analysis; Test Interpretation; *Test Reliability
IDENTIFIERS *Domain Referenced Tests; Kappa Coefficient; Loss Function; Phi Coefficient

ABSTRACT

A large number of seemingly diverse coefficients have been proposed as indices of dependability, or reliability, for domain-referenced and/or mastery tests. In this paper, it is shown that most of these indices are special cases of two generalized indices of agreement: one that is corrected for chance, and one that is not. The special cases of these two indices are determined by assumptions about the nature of the agreement function or, equivalently, the nature of the loss function for the testing procedure. For example, indices discussed by Huynh, Subkoviak, and Swaminathan, Hambleton, and Algina employ a threshold agreement, or loss, function whereas, indices discussed by Brennan and Kane and Livingston employ a squared error loss function. Since all of these indices are discussed within a single general framework, the differences among them in their assumptions, properties, and uses can be exhibited clearly. For purposes of comparison, norm-referenced generalizability coefficients are also developed and discussed within this general framework. (Author/CTM)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED185076

ACT TECHNICAL BULLETIN NO. 28

# Agreement Coefficients as Indices of Dependability for Domain-Referenced Tests

by

Michael T. Kane

National League for Nursing

and

Robert L. Brennan

American College Testing Program

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
R. Ferguson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The Research and Development Division

The American College Testing Program

P. O. Box 168, Iowa City, Iowa 52240

December, 1977

TM800 063

## Abstract

A large number of seemingly diverse coefficients have been proposed as indices of dependability, or reliability, for domain-referenced and/or mastery tests. In this paper, it is shown that most of these indices are special cases of two generalized indices of agreement, one that is corrected for chance, and one that is not. The special cases of these two indices are determined by assumptions about the nature of the agreement function or, equivalently, the nature of the loss function for the testing procedure. For example, indices discussed by Huynh (1976), Subkoviak (1976), and Swaminathan, Hambleton, and Algina (1974) employ a threshold agreement, or loss, function; whereas, indices discussed by Brennan and Kane (1977a, 1977b) and Livingston (1972a) employ a squared error loss function. Since all of these indices are discussed within a single general framework, the differences among them in their assumptions, properties, and uses can be exhibited clearly. For purposes of comparison, norm-referenced generalizability coefficients are also developed and discussed within this general framework.

# Table of Contents


| | Page |
|---|---|
| Glossary of Symbols | iii |
| Introduction | 2 |
| A Pair of General Indices of Dependability | 5 |
| Agreement Function | 5 |
| Maximum Agreement and the Index $\theta$ | 6 |
| Chance Agreement and the Index $\theta_c$ | 9 |
| Loss | 11 |
| Interpretation of $\theta$ and $\theta_c$ | 12 |
| $\theta$ and $\theta_c$ for Threshold Agreement | 13 |
| Threshold Agreement Function | 13 |
| The Index $\theta(t)$ | 14 |
| The Index $\theta_c(t)$ | 15 |
| Threshold Loss | 16 |
| $\theta$ and $\theta_c$ for Domain-Referenced Agreement | 18 |
| Domain-Referenced Agreement Function | 20 |
| The Index $\theta(d)$ | 21 |
| The Index $\theta_c(d)$ | 23 |
| Domain-Referenced Loss and $\sigma^2(\Delta)$ | 24 |
| Interpretation of $\theta(d)$ and $\theta_c(d)$ | 26 |
| Domain-Referenced Agreement without a Cutting Score | 27 |
| $\theta$ and $\theta_c$ for Norm-Referenced Agreement | 29 |
| The Indices $\theta(g)$ and $\theta_c(g)$ | 30 |
| Norm-Referenced Loss and $\sigma^2(\delta)$ | 31 |
| Interpretation of $\theta(g) = \theta_c(g)$ | 33 |

Table of Contents (Continued)


| | Page |
|---|---|
| The Effect of Item Sampling on the Indices $\theta$ and $\theta_c$ | 34 |
| Items Nested Within Persons in the D Study | 34 |
| Implications for Norm-Referenced and Domain-Referenced Indices of Dependability | 36 |
| Summary and Conclusions | 39 |
| Choosing an Index of Dependability | 39 |
| Prior Information | 41 |
| Assumptions about Parallel Tests | 42 |
| Concluding Comments | 44 |
| Reference Notes | 45 |
| References | 46 |
| Footnote | 50 |
| Table 1 | 51 |

## Glossary of Symbols

| Symbol | Definition |
|---|---|
| $A$ | Expected agreement |
| $A_c$ | Chance agreement |
| $A_m$ | Maximum expected agreement |
| $I$, $J$ | Particular samples of $n$ items |
| $L$ | Loss or expected disagreement |
| $S$ or $X$ | Score random variable |
| $a$ | Any agreement function |
| $d$ | Domain-referenced agreement function |
| $e$ | Residual error |
| $g$ | Norm-referenced agreement function |
| $i$, $j$ | Subscripts for score categories |
| $k$, $l$ | Test subscripts |
| $n$ | Usually, the number of items in a test (or instance of a testing procedure); $n + 1$ is always the total number of score categories. |
| $p$ or $Pr$ | Probability |
| $s$ | Realization of score random variable S |
| $t$ | Threshold agreement function |
| $v$, $w$ | Person subscripts |
| $\beta$ | Item effect |
| $\delta$, $\Delta$ | Different types of error |
| $\theta$ | Index of dependability not corrected for chance |

Glossary of Symbols (Continued)

| Symbol | Definition |
|---|---|
| $\theta_c$ | Index of dependability corrected for chance |
| $\lambda$ | Cutting score |
| $\mu$ | Grand mean in population and universe |
| $\pi$ | Person effect |
| $\Sigma$ | Summation |
| $\mathcal{E}$ | Expected value |
| $\sigma^2$ | Variance component |

## Introduction

Glaser and Nitko (1971) define a criterion-referenced test as "one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards" (p. 653). This is probably the best-known definition of a criterion-referenced test, but others have been proposed (e.g., Ivens, 1970; Kriewall, 1969; and Livingston, 1972a). Nothing in the Glaser and Nitko definition, or in most other definitions of "criterion-referenced test," necessitates the existence or use of a single criterion or cutting score as a "specified performance standard." However, much of the literature subsumed under the heading of criterion-referenced measurement does, in fact, postulate the existence of a single cutting score. Since this inconsistency in terminology can lead to confusion, we prefer to reserve the term mastery test for a criterion-referenced test with a single fixed mastery cutting score.

Hively (1974) and Millman (1974), among others, suggest using the descriptor "domain-referenced test" rather than "criterion-referenced test." They note that the word "criterion" is ambiguous in some contexts, and they argue that the word "domain" provides a more direct specification of the entire set of items or tasks under consideration. If one accepts these arguments, a mastery test can be defined as a domain-referenced test with a single cutting score.

One can also distinguish between a particular type of test (g.e., norm-referenced or domain-referenced) and the scores (or interpretation of scores) resulting from a test. For example, the scores from any test might be given norm-referenced or domain-referenced interpretations. Indeed, most of the literature that treats issues of dependability (or reliability) of domain-referenced tests actually treats the dependability of a particular set of scores that are given (or provide) a domain-referenced or mastery interpretation. In this paper, to obviate verbal complexity, we will often refer to norm-referenced, domain-referenced, and mastery "tests"; however, a more complete verbal description would refer to scores that are given (or provide) norm-referenced, domain-referenced, or mastery interpretations for a particular testing procedure.

Since Popham and Husek (1969) challenged the appropriateness of correlation coefficients as indices of reliability for domain-referenced and mastery tests, considerable effort has been devoted to developing more appropriate indices. Most of these indices have been proposed as measures of reliability; however, we prefer to use the more generic term, dependability, in order to avoid unwarranted associations with the classical theory of reliability for norm-referenced tests.

Since a large number of seemingly diverse coefficients have been proposed, it has been difficult for evaluators to distinguish among them in meaningful ways. In this paper, we show that most of these indices can be classified into two broad categories depending on their underlying (and sometimes unstated) assumptions about the nature of agreement or, equivalently, the nature of loss in the testing procedure. For example, indices discussed by

Carver (1970), Huynh (1976), Marshall and Haertel (Note 1), Subkoviak (1976), and Swaminathan, Hambleton, and Algina (1974) employ a threshold agreement, or loss, function; whereas, indices discussed by Brennan and Kane (1977a, 1977b) and Livingston (1972a, 1972b, 1972c, 1973) employ a squared-error loss function. We will also show that within these two broad categories indices can be differentiated with respect to whether or not they incorporate a correction for chance agreement. In addition, we will examine both the nature of agreement and the role of chance agreement in norm-referenced testing.

We begin by using notions of agreement in order to develop two generalized indices of dependability. One of these indices is corrected for chance agreement, and the other is not. For both of these general indices no specific agreement function is assumed. The actual indices that result from several specific agreement functions are then examined in detail. This examination of a large number of indices of dependability, within a single consistent framework, makes it possible to compare and contrast the assumptions, properties, interpretations, and uses of these indices.

## A Pair of General Indices of Dependability

### Agreement Function

A general expression for the dependability of a testing procedure can be derived by examining the expected agreement between two randomly selected instances of the testing procedure. Any particular instance of a testing procedure will be referred to as a "test." No assumptions need to be made about the nature of the tests, the details of their administration, or their scoring. Since the instances, or tests, are randomly selected from a universe of possible instances, they are randomly parallel. Therefore, the expected distribution of outcomes for the population is assumed to be the same for all instances of the testing procedure. This does not imply that the distributions of outcomes are necessarily identical for all tests; that is, we are not making the stronger assumption of classically parallel tests.

The degree of agreement between any two scores, $\underline{s}_i$ and $\underline{s}_j$, is defined by an agreement function, $\underline{a}(\underline{s}_i, \underline{s}_j)$. The scores, $\underline{s}_i$ and $\underline{s}_j$, may be raw scores, or they may be transformed in some way. For convenience, we shall assume that only a finite number of scores ($\underline{s}_o, \ldots \underline{s}_i, \ldots \underline{s}_n$), may result from the use of the testing procedure. The form of the agreement function defines what is meant by agreement in any particular context. In general, the agreement function will reflect intuitive and, therefore, somewhat arbitrary notions of the relative degree of agreement for different pairs of scores. As we shall see, the choice of an agreement function implies the choice of a loss function, where loss is defined as the difference between maximum possible agreement and observed agreement in a particular context.

Although we shall not assume any particular form for an agreement function in our development of a general index of dependability, it is reasonable to impose some conditions on the class of functions that will be accepted as agreement functions. In the discussion that follows, it is assumed that all agreement functions satisfy the following three conditions:

$$\begin{aligned} &\text{(i)} \quad a(s_i, s_i) \geq 0; \\ &\text{(ii)} \quad a(s_i, s_j) = a(s_j, s_i); \text{ and} \\ &\text{(iii)} \quad a(s_i, s_i) + a(s_j, s_j) \geq 2a(s_i, s_j). \end{aligned} \tag{1}$$

Given that we are examining the agreement between randomly parallel tests, the first two of these conditions are certainly natural. The third condition simply states that the agreement assigned to any pair of scores, $s_i$ and $s_j$, cannot be greater than the average of the agreements that result from pairing each of these scores with itself. All the agreement functions discussed in this paper satisfy these three conditions.

Maximum Agreement and the Index $\theta$

The score for person $v$ on the $k$-th instance of the testing procedure can be represented by the random variable $S_{vk}$. Similarly, $S_{wl}$ is the score for person $w$ on test $l$. For every person $v$ and every test $k$, $S_{vk}$ takes one of the values $s_0, \ldots . s_n$. We might, then, take as our index of dependability the expected agreement given by:

$$A = \mathop{\mathcal{E}}_{v,k,l} a(S_{vk}, S_{vl}), \tag{2}$$

where expectation is taken over the population of persons and over pairs of tests that are independently sampled from the universe of tests. The expected agreement may also be represented in terms of the joint distribution of scores on the two tests:

$$\underline{A} = \sum_{\underline{i},\underline{j}=0}^{n} \underline{a}(\underline{s}_{\underline{i}}, \underline{s}_{\underline{j}}) \cdot \underline{Pr}(\underline{S}_{\underline{vk}} = \underline{s}_{\underline{i}}, \underline{S}_{\underline{vl}} = \underline{s}_{\underline{j}}) \qquad (3)$$

where $\underline{Pr}(\underline{S}_{\underline{vk}} = \underline{s}_{\underline{i}}, \underline{S}_{\underline{vl}} = \underline{s}_{\underline{j}})$ is the probability that a randomly chosen person, $\underline{v}$, got scores $\underline{s}_{\underline{i}}$ and $\underline{s}_{\underline{j}}$, on randomly chosen tests, $\underline{k}$ and $\underline{l}$. Equations 2 and 3 represent the same quantities expressed in two different ways. In the following discussion, we shall use whichever of these expressions is most convenient for the issue under consideration. The notation in Equation 3 can be simplified by letting

$$\underline{a}_{\underline{ij}} = \underline{a}(\underline{s}_{\underline{i}}, \underline{s}_{\underline{j}})$$

and

$$\underline{p}_{\underline{ij}} = \underline{Pr}(\underline{S}_{\underline{vk}} = \underline{s}_{\underline{i}}, \underline{S}_{\underline{vl}} = \underline{s}_{\underline{j}}).$$

Equation 3 can then be written as

$$\underline{A} = \sum_{\underline{i},\underline{j}=0}^{n} \underline{a}_{\underline{ij}}\underline{p}_{\underline{ij}}. \qquad (4)$$

However, the index, $\underline{A}$, depends on the scale chosen for $\underline{a}_{\underline{ij}}$, and can be made arbitrarily large by multiplying $\underline{a}_{\underline{ij}}$ by a sufficiently large constant. One way to correct this problem is to take as the index of agreement:

$$\theta = \frac{A}{A_m} . \tag{5}$$

In Equation 5, $A_m$ is the expected agreement between an instance of the testing procedure and itself:

$$A_m = \sum_{v,k} a(S_{vk}, S_{vk}) \tag{6}$$

$$= \sum_{i=0}^{n} a(s_i, s_i) \cdot Pr(S_{vk} = s_i);$$

and in the simpler notation,

$$A_m = \sum_{i=0}^{n} a_{ii} p_i, \tag{7}$$

where $p_i$ is the probability that a randomly selected person will get the score, $s_i$, on a randomly chosen instance of the testing procedure. $A$ is equal to $A_m$ when every person in the population gets the same score on every instance of the testing procedure; i.e., when all instances of the testing procedure are in perfect agreement in the assignment of scores, $s_i$, to persons in the population.

Using the three conditions in Equation 1, it is easy to show that for any marginal distribution, $A_m$ is the maximum value of $A$.

Since $p_i$ is a marginal probability,

$$A_m = \sum_i a_{ii} p_i = \sum_{i,j} a_{ii} p_{ij},$$

and

$$A_m = \sum_j a_{jj} p_j = \sum_{i,j} a_{jj} p_{ij}.$$

Therefore, $A_m$ can be written as:

$$A_m = \sum_{i,j} \left[ \frac{a_{ii} + a_{jj}}{2} \right] p_{ij}.$$

Now, using Assumption iii in Equation 1

$$A_m \geq \sum_{i,j} a_{ij} p_{ij};$$

and using Equation 4

$$A_m \geq A.$$

From the definition of $\theta$ in Equation 5, therefore, it follows that $A_m$ is the maximum value of the agreement function, and $\theta$ is less than or equal to one.

Chance Agreement and the Index $\theta_c$

The coefficient $\theta$ provides a general index of dependability for any agreement function, but it does not consider the contribution of chance agreement to the dependability of measurement. As we shall see, $\theta$ may be large even when scores are randomly assigned to persons on each instance of the testing procedure. When we say that a score is assigned to examinee $v$, by chance, we mean that the score is randomly selected from the distribution of scores for the population of persons on the universe of tests. The assignment of $s_i$ to examinee $v$, by chance, depends only on the marginal probability, $p_i$, of the score $s_i$, and not on the examinee's performance. Therefore, for chance assignment, the score assigned to an examinee on any particular instance of the testing procedure is independent on the score assigned on any other instance.

The contribution of chance agreement can be examined by taking the expected agreement between the score, $S_{vk}$, for person $v$ on the $k$-th test, and the score, $S_{wl}$, for an independently sampled person, $w$, on an independently sampled test, $l$:

$$A_c = \sum_{v,w,k,l} a(S_{vk}, S_{wl}) \tag{8}$$

or

$$A_c = \sum_{i,j=0}^{n} a(s_i, s_j) \cdot Pr(S_{vk} = s_i, S_{wl} = s_j). \tag{9}$$

Since both persons, $v$ and $w$, and tests, $k$ and $l$, are sampled independently,

$$Pr(S_{vk} = s_i, S_{wl} = s_j) = Pr(S_{vk} = s_i) \cdot Pr(S_{wl} = s_j), \tag{10}$$

where $Pr(S_{vk} = s_i)$ is the marginal probability that a randomly selected person will get the score, $s_i$, on a randomly chosen test. Substituting Equation 10 in Equation 9, and using the simplified notation introduced earlier, we have

$$A_c = \sum_{i,j=0}^{n} a_{ij} p_i p_j. \tag{11}$$

For any agreement function, Equation 11 depends only on the marginal distribution for a single administration of the testing procedure. $A_c$ is the expected agreement for pairs of scores when each score is independently sampled from the marginal distribution of the population.

A general index of dependability, corrected for chance, can then be defined as:

$$\theta_{\underline{C}} = \frac{\underline{A} - \underline{A}_{\underline{C}}}{\underline{A}_{\underline{m}} - \underline{A}_{\underline{C}}} . \qquad (12)$$

The numerator of Equation 12 provides a measure of how much the expected agreement for the testing procedure exceeds the expected agreement due to chance. Since

$$\underline{A}_{\underline{m}} \geq \underline{A},$$

$$\underline{A}_{\underline{m}} - \underline{A}_{\underline{C}} \geq \underline{A} - \underline{A}_{\underline{C}},$$

the denominator of Equation 12 is the maximum value of the numerator, and $\theta_{\underline{C}}$ is less than or equal to one.

Loss

Although most of the discussion in this paper is concerned with agreement functions, it will be useful in some places to discuss the expected disagreement or loss associated with testing procedures. The expected loss, $\underline{L}$, for any testing procedure is defined as the difference between the maximum expected agreement and the expected agreement:

$$\underline{L} = \underline{A}_{\underline{m}} - \underline{A} . \qquad (13)$$

Using this definition, Equation 5 can be written:

$$\theta = \frac{\underline{A}}{\underline{A} + \underline{L}} ; \qquad (14)$$

and Equation 12 can be written:

$$\theta_{\underline{C}} = \frac{\underline{A} - \underline{A_C}}{(\underline{A} - \underline{A_C}) + \underline{L}} . \qquad (15)$$

Note that Equations 14 and 15 have the form of classical reliability coefficients, or generalizability coefficients, with $\underline{L}$ taking the place of error variance (see Cronbach, Gleser, Nanda, and Rajaratnam, 1972).

Interpretation of $\theta$ and $\theta_{\underline{C}}$

The two indices, $\theta$ and $\theta_{\underline{C}}$, address different questions about dependability. Some of the properties of these indices will be discussed more fully later in the context of particular agreement functions, but a brief statement is appropriate here.

$\theta$ indicates how closely, in terms of the agreement function, the scores for any examinee can be expected to agree. $\theta_{\underline{C}}$ indicates how closely (again, in terms of the agreement function) the two scores for an examinee can be expected to agree, with the contribution of chance agreement removed. For the agreement functions discussed in this paper, $\theta_{\underline{C}}$ is less than or equal to $\theta$.

The index, $\theta$, therefore, characterizes the dependability of decisions, or estimates, based on the testing procedure. The magnitude of $\theta$ depends, in part, on chance agreement; it may be greater than zero even when decisions based on the testing procedure are no more dependable than decisions based on marginal probabilities in the population. The index, $\theta_{\underline{C}}$, characterizes the contribution of the testing procedure to the dependability of the decisions, over what would be expected on the basis of chance agreement. $\theta$ provides an estimate of the dependability of the decisions based on the testing procedure;

$\theta_c$ provides an estimate of the contribution of the testing procedure to the dependability of such decisions. The two indices provide answers to different questions. The issue is not which of these indices is best, but rather which is appropriate in a given context.

## $\theta$ and $\theta_c$ for Threshold Agreement

### Threshold Agreement Function

One common use of tests is to classify examinees into two or more mutually exclusive categories. If there are only two categories, or if the categories are unordered, then a plausible agreement function for the classification procedure is given by the threshold agreement function, $t$:

$$t(S_{vk}, S_{wl}) = \begin{cases} 1 \text{ if } S_{vk} = S_{wl} \\ 0 \text{ if } S_{vk} \neq S_{wl} \end{cases}, \qquad (16)$$

where $S_{vk}$ is the score (in this case the category) for examinee $v$ on the test $k$. Equation 16 can be expressed more succinctly as:

$$t_{ij} = t(s_i, s_j) = \begin{cases} 1 \text{ if } s_i = s_j \\ 0 \text{ if } s_i \neq s_j \end{cases}, \qquad (17)$$

where the score $s_i$ represents assignment to the $i$-th category. Equation 17 has the advantage of notational simplicity, whereas Equation 16 is a more detailed statement of the threshold agreement function, $t$. For either expression, the

assigned agreement is one if the examinee is placed in the same category on both administrations of the procedure, and agreement is zero if the examinee is placed in different categories on the two administrations. It is easily verified that the agreement function in Equation 16 satisfies the three conditions in Equation 1.

The Index $\theta(\underline{t})$

Substituting $\underline{t}_{ij}$, given by Equation 17, for $\underline{a}_{ij}$ in Equation 4, we obtain the expected agreement for classification procedures:

$$\underline{A}(\underline{t}) = \sum_{\underline{i},\underline{j}} \underline{t}_{ij}\underline{p}_{ij} = \sum_{\underline{i}} \underline{p}_{ii}. \tag{18}$$

The maximum agreement is given by:

$$\underline{A}_{\underline{m}}(\underline{t}) = \sum_{\underline{i}} \underline{t}_{ii}\underline{p}_{\underline{i}} = 1. \tag{19}$$

The definition of $\theta$, an index of dependability not corrected for chance, is provided in Equation 5. For the threshold agreement function, this index is given by:

$$\theta(\underline{t}) = \frac{\underline{A}(\underline{t})}{\underline{A}_{\underline{m}}(\underline{t})}, \tag{20}$$

and substituting Equations 18 and 19 in Equation 20, we obtain

$$\theta(\underline{t}) = \sum_{\underline{i}} \underline{p}_{ii}. \tag{21}$$

Equation 21 states that the dependability of the classification procedure is simply the probability that a randomly chosen examinee will be placed in the same category on two randomly chosen instances of the procedure. Note that $\underline{A}(\underline{t})$, $\underline{A}_{\underline{m}}(\underline{t})$ and $\theta(\underline{t})$ are all equal to one if the classification procedure consistently places all examinees into a single category.

Equation 21 is stated in terms of population parameters. Estimates of $\theta(\underline{t})$, based on two administrations of the testing procedure have been discussed by Berger (Note 2), Carver (1970), and Swaminathan, et al. (1974). Estimates of $\theta(\underline{t})$, based on a single administration of the testing procedure, have been discussed by Marshall and Haertel (Note 1), Subkoviak (1976, Note 3), and by Subkoviak and Albrecht (Note 4).

The Index $\theta_{\underline{c}}(\underline{t})$

The definition of $\theta_{\underline{c}}$, an index of dependability corrected for chance, is provided in Equation 12. For the threshold agreement function in Equation 17, the expected agreement due to chance is:

$$\underline{A}_{\underline{c}}(\underline{t}) = \sum_{\underline{i},\underline{j}} \underline{t}_{\underline{i}\underline{j}} \underline{p}_{\underline{i}} \underline{p}_{\underline{j}}$$

$$= \sum_{\underline{i}} \underline{p}_{\underline{i}}^{2} . \qquad (22)$$

Subtracting $\underline{A}_{\underline{c}}(\underline{t})$ from the numerator and denominator of $\theta(t)$ in Equation 20, we have the index of dependability corrected for chance, for a threshold agreement function:

$$\theta_{\underline{c}}(\underline{t}) = \frac{\underline{A}(\underline{t}) - \underline{A}_{\underline{c}}(\underline{t})}{\underline{A}_{\underline{m}}(\underline{t}) - \underline{A}_{\underline{c}}(\underline{t})}$$

$$= \frac{\Sigma \underline{p}_{\underline{i}\underline{i}} - \Sigma \underline{p}_{\underline{i}}^{2}}{1 - \Sigma \underline{p}_{\underline{i}}^{2}} . \qquad (23)$$

In the special case where all examinees are consistently placed in a single category, $A_c(t)$ is equal to one and $\theta_c(t)$ is indeterminate.

The index $\theta_c(t)$ in Equation 23 is identical to Cohen's (1960) coefficient kappa, and to Scott's (1955) coefficient, under our assumption that the expected marginal distributions for the two instances of the testing procedure are identical. As such, $\theta_c(t)$ has been proposed by Huynh (1976) and Swaminathan et al (1974) as an index of reliability for mastery tests with a single cutting score.

Threshold Loss

The loss associated with a threshold agreement function can be determined by subtracting Equation 18 from Equation 19:

$$L(t) = A_m(t) - A(t)$$

$$= \sum_{i \neq j} p_{ij}.$$

If the two instances of the testing procedure assign the person to the same category, the loss is zero. If the two instances assign a person to different categories, the loss is one, regardless of which categories are involved. This is consistent with the usual definition of a threshold loss function (see Hambleton and Novick, 1973).

Interpretation of $\theta(t)$ and $\theta_c(t)$

The first block of Table 1 summarizes results for the parameters, $A(t)$, $A_m(t)$, $A_c(t)$, and $L(t)$, and the agreement indices, $\theta(t)$ and $\theta_c(t)$, for the threshold agreement function, $t$.

---

Insert Table 1 about here

---

As noted earlier, $\theta(\underline{t})$ will be equal to one whenever all instances of the categorization procedure place all persons into one category. The testing procedure used to assign persons to categories is then perfectly dependable and completely superfluous. Once it is established that all, or almost all, persons fall into one category, there is little to be gained by administering tests.

If almost everyone is in one category, the expected chance agreement, $A_c(\underline{t})$, will be close to $A_m(\underline{t})$, the maximum expected agreement. Under these circumstances, it would be difficult for any testing procedure to provide a significant improvement in dependability over chance assignment. Consequently, the coefficient corrected for chance, $\theta_c(\underline{t})$, will tend to be small whenever the testing procedure places almost everyone in the same category.

Therefore, $\theta_c(\underline{t})$ is liable to one of the objections raised by Popham and Husek (1969) against classical reliability coefficients as indices for mastery tests--namely, $\theta_c(\underline{t})$ may be close to zero even when individuals are consistently placed in the correct category. However, this property of $\theta_c(\underline{t})$ does not point to any basic flaw in the coefficient, but only to a possible misinterpretation of the coefficient. A low value of $\theta_c(\underline{t})$ does not necessarily indicate that assignments to categories are inconsistent from one administration to the next. Rather, a low value of $\theta_c(\underline{t})$ indicates that the use of the testing procedure in classifying individuals is not much more dependable than a process of random assignment based on prior information about the population (i.e., the marginals in the population). Note

that $\theta(\underline{t})$ is large whenever the classification of examinees is consistent from one instance of the testing procedure to another; therefore, $\theta(\underline{t})$ is not subject to Popham and Husek's objection.

Contrary to a suggestion by Subkoviak (1976), the two coefficients developed from the threshold agreement function are not appropriate when there are more than two categories, and these categories are ordered in some way. The threshold agreement function in Equation 16 assumes that the categories are not ordered in any way.

## $\theta$ and $\theta_c$ for Domain-Referenced Agreement

In our discussion of domain-referenced agreement, we shall assume that, for each instance of the testing procedure, a random sample of $\underline{n}$ items is drawn from some infinite domain (or universe) of items, and the sample of items is administered to all examinees.

In the last section we used a threshold loss function to examine the dependability of procedures that assign each examinee to one of a set of qualitative categories. In this section, we shall examine the dependability of domain-referenced testing procedures. We shall emphasize the use of such procedures for mastery decisions with a single cutting score, but we shall also discuss the use of domain-referenced tests in the absence of a specified cutting score.

The score for person $\underline{v}$ on item $\underline{i}$ can be represented by a general linear model:

$$X_{\underline{vi}} = \mu + \pi_{\underline{v}} + \beta_{\underline{i}} + (\pi\beta,\underline{e})_{\underline{vi}} \qquad (24)$$

where

$X_{vi}$ = observed score for person $v$ on item $i$;

$\mu$ = grand mean in the population of persons and the universe of items;

$\pi_v$ = effect for person $v$;

$\beta_i$ = effect for item $i$;

$(\pi\beta,e)_{vi}$ = effect for the interaction of person $v$ and item $i$, which is confounded with residual error;

and all effects are assumed to be independent random effects. In the usual case, where each examinee responds once to each item, the interaction effect and the residual error are completely confounded and, therefore, these two effects are combined in Equation 24.

In the discussion that follows, the observed score for person $v$ will be taken to be the mean score over the sample of $n$ items. To be consistent with our earlier notation, we will let the subscript $I$ indicate a particular sample of $n$ items, and we will designate a person's observed mean score as:

$$S_{vI} = \mu + \pi_v + \beta_I + (\pi\beta,e)_{vI} . \qquad (25)$$

Similarly, the score for person, $w$, on the $J$-th sample of $n$ items is

$$S_{wJ} = \mu + \pi_w + \beta_J + (\pi\beta,e)_{wJ} . \qquad (26)$$

Note that $S_{vI}$ and $S_{wJ}$ are observed scores; they are not the same as $S_{vk}$ and $S_{wl}$ used previously to denote categories to which persons are assigned.

Domain-Referenced Agreement Function

One source of difficulty with $\theta(\underline{t})$ and $\theta_{\underline{c}}(\underline{t})$ for mastery testing is the nature of the threshold agreement function. When mastery testing is used to make placement decisions, errors may involve very different degrees of loss. If a mastery test consisting of a sample from a universe of spelling words has a cut-off of 80%, the consequences of misclassifying a student with a universe score of 79% are likely to be far less serious than the consequences of misclassifying a student with a universe score of 40%. A threshold loss function assigns the same loss to both of these cases.

This suggests that the agreement function for domain-referenced tests that are used for mastery decisions should involve the distance of the observed score from the cutting score. For a cutting score, $\lambda$, the domain-referenced agreement function is defined by:

$$\underline{d}(S_{\underline{vI}}, S_{\underline{wJ}}) = (S_{\underline{vI}} - \lambda)(S_{\underline{wJ}} - \lambda) , \qquad (27)$$

where $\underline{I}$ and $\underline{J}$ refer to independent samples of n items. Equation 27 assigns a positive agreement to two scores that result in the same classification, mastery or non-mastery. It assigns a negative agreement to two scores that result in different classifications. In either case, the magnitude of the agreement depends on the magnitudes of two deviation scores, $(S_{\underline{vI}} - \lambda)$ and $(S_{\underline{wJ}} - \lambda)$. If both of these deviation scores are close to zero, indicating a "borderline case," the magnitude of the agreement function will be close to zero. If both of these deviation scores are large and in the same direction, indicating strong agreement, the domain-referenced agreement function will be large and positive. If both deviations are large and in opposite directions, indicating strong disagreement, the domain-referenced agreement function will be large and negative.

The domain-referenced agreement function in Equation 27 is similar to the definition of agreement used by Livingston (1972a) in developing an index of reliability for mastery tests. However, Livingston assumed that the two tests were parallel in the sense of classical test theory. We base our analysis on generalizability theory which makes the weaker assumption that the tests are randomly parallel. As a result, the indices derived here differ from Livingston's coefficient in several significant ways.

The Index $\theta(\underline{d})$

Using the domain-referenced agreement function in Equation 27 and the definition of expected agreement in Equation 2, we obtain

$$\underline{A}(\underline{d}) = \underset{\underline{v},\underline{I},\underline{J}}{\xi} [(\underline{S}_{\underline{v}\underline{I}} - \lambda)(\underline{S}_{\underline{v}\underline{J}} - \lambda)]. \qquad (28)$$

Now, using Equation 25 to replace $\underline{S}_{\underline{v}\underline{I}}$ and $\underline{S}_{\underline{v}\underline{J}}$ in Equation 28,

$$\underline{A}(\underline{d}) = \underset{\underline{v},\underline{I},\underline{J}}{\xi} [(\mu - \lambda) + \pi_{\underline{v}} + \beta_{\underline{I}} + (\pi\beta,\underline{e})_{\underline{v}\underline{I}}] \cdot [(\mu - \lambda) + \pi_{\underline{v}} + \beta_{\underline{J}} + (\pi\beta,\underline{e})_{\underline{v}\underline{J}}]. \qquad (29)$$

Since the effects $\pi$, $\beta$, and $(\pi\beta,\underline{e})$ are assumed to be sampled independently, and $\mu$ and $\lambda$ are constants, the expected value of the cross-products are zero; and Equation 29 reduces to

$$\underline{A}(\underline{d}) = \xi (\mu - \lambda)^2 + \xi \pi_{\underline{v}}^2 + \xi \beta_{\underline{I}}\beta_{\underline{J}} + \xi (\pi\beta,\underline{e})_{\underline{v}\underline{I}}(\pi\beta,\underline{e})_{\underline{v}\underline{J}}. \qquad (30)$$

Because the two sets of items are independently sampled, the last two terms in Equation 30 equal zero. Also, by the definition of a variance component $\sigma^2(\pi) = \sum_{\underline{v}} \pi_{\underline{v}}^2$; and, therefore, the expected agreement, for the domain-referenced agreement function, is

$$\underline{A}(\underline{d}) = (\mu - \lambda)^2 + \sigma^2(\pi). \tag{31}$$

Similarly, the maximum expected agreement is found by using Equation 27 and the definition of maximum expected agreement in Equation 6:

$$\underline{A}_{\underline{m}}(\underline{d}) = \sum_{\underline{v},\underline{I}} (\underline{S}_{\underline{v}\underline{I}} - \lambda)^2$$

$$= (\mu - \lambda)^2 + \sigma^2(\pi) + \frac{\sigma^2(\beta)}{\underline{n}} + \frac{\sigma^2(\pi\beta,\underline{e})}{\underline{n}}, \tag{32}$$

where $\underline{n}$ is the number of items sampled for each instance of the testing procedure. Substituting Equations 31 and 32 in Equation 5, the index of dependability for mastery decisions is given by:

$$\theta(\underline{d}) = \frac{(\mu - \lambda)^2 + \sigma^2(\pi)}{(\mu - \lambda)^2 + \sigma^2(\pi) + \frac{\sigma^2(\beta)}{\underline{n}} + \frac{\sigma^2(\pi\beta,\underline{e})}{\underline{n}}}. \tag{33}$$

Equations for estimating $\theta(\underline{d})$ have been discussed by Brennan and Kane (1977a). The constant, $\underline{n}$, appears in Equations 32 and 33 because the observed scores are assumed to be averages over $\underline{n}$ items.

It is clear from Equation 33 that $\theta(\underline{d})$ will tend to be large when $(\mu - \lambda)^2$ is large (i.e., when the population mean is very different from the cutting score) even if $\sigma^2(\pi)$ is zero. If all examinees have the same universe score, $\sigma^2(\pi)$ is zero, and $(\mu - \lambda)^2$ provides a measure of the strength of the signal that needs to be detected for accurate classification (see Brennan and Kane, 1977b). If this signal is large the required decisions are easy to make, and it is possible in such cases to classify examinees dependably, even if the test being used does not provide dependable information about individual differences among universe scores.

The Index $\theta_{\underline{c}}(\underline{d})$

Using the domain-referenced agreement function in Equation 27 and the definition of chance agreement in Equation 8, the expected agreement due to chance is:

$$A_{\underline{c}}(\underline{d}) = \mathop{E}_{\underline{v},\underline{w},\underline{I},\underline{J}} [(S_{\underline{v}\underline{I}} - \lambda)(S_{\underline{w}\underline{J}} - \lambda)] \quad .$$

Replacing $S_{\underline{v}\underline{I}}$ and $S_{\underline{w}\underline{J}}$ from Equations 25 and 26, and taking the expected value over $\underline{v}$, $\underline{w}$, $\underline{I}$, and $\underline{J}$, the expected chance agreement for the domain-referenced agreement function is:

$$A_{\underline{c}}(\underline{d}) = (\mu - \lambda)^2 \quad . \tag{34}$$

Subtracting $A_{\underline{c}}(\underline{d})$ from the numerator and denominator of Equation 33, the domain-referenced index of dependability, corrected for chance agreement is:

$$\theta_{\underline{c}}(\underline{d}) = \frac{\sigma^2(\pi)}{\sigma^2(\pi) + \dfrac{\sigma^2(\beta)}{\underline{n}} + \dfrac{\sigma^2(\pi\beta,\underline{e})}{\underline{n}}} \quad . \tag{35}$$

The estimation of this index is discussed by Brennan and Kane (1977b), and its relationship to KR-21 is discussed by Brennan (1977b).

Note that $\theta_{\underline{c}}(\underline{d})$ is zero when $\sigma^2(\pi)$ is zero. If the test is to provide more dependable classification of examinees than could be achieved by chance, it must differentiate among the examinees. Therefore, some variability in universe scores is required if the test is to make a contribution to the dependability of the decision procedure.

Domain-Referenced Loss and $\sigma^2(\Delta)$

For the domain-referenced agreement function, the expected loss can be found by subtracting Equation 31 from Equation 32:

$$\underline{L}(\underline{d}) = \underline{A}_{\underline{m}}(\underline{d}) - \underline{A}(\underline{d})$$

$$= \frac{\sigma^2(\beta)}{\underline{n}} + \frac{\sigma^2(\pi\beta,\underline{e})}{\underline{n}} . \qquad (36)$$

The loss, $\underline{L}(\underline{d})$, is therefore equal to the error $\sigma^2(\Delta)$, which is discussed by Cronbach et al (1972), Brennan (1977a, 1977b) and Brennan and Kane (1977a, 1977b). The error variance $\sigma^2(\Delta)$ is appropriate for domain-referenced testing, in general, and for mastery testing, in particular.

In mastery testing, we are interested in "the degree to which the student has attained criterion performance" (Glaser, 1963, p. 519), independent of the performance of other students. That is, we are not primarily interested in the relative ordering of examinees' universe scores; rather, we are interested in the difference between each examinee's universe score and the absolute standard defined by the mastery cutting score. In generalizability theory, the universe score for examinee $\underline{v}$ is, by definition,

$$\mu_{\underline{v}} = \mathop{\mathcal{E}}_{\underline{I}} S_{\underline{vI}} = \mu + \pi_{\underline{v}},$$

where $S_{vI}$ is defined by Equation 25, and the expectation is taken over all possible random samples of $\underline{n}$ items from the universe of items. Therefore, for a mastery test, the error for examinee $\underline{v}$ is:

$$\begin{aligned}
\Delta_{\underline{v}} &= (S_{\underline{vI}} - \lambda) - (\mu_{\underline{v}} - \lambda) \\
&= S_{\underline{vI}} - \mu_{\underline{v}} \\
&= [\mu + \pi_{\underline{v}} + \beta_{\underline{I}} + (\pi\beta,\underline{e})_{\underline{vI}}] - [\mu + \pi_{\underline{v}}] \\
&= \beta_{\underline{I}} + (\pi\beta,\underline{e})_{\underline{vI}};
\end{aligned}$$

and the variance of $\Delta_{\underline{v}}$ over persons and random samples of $\underline{n}$ items is $\sigma^2(\Delta)$, given by Equation 36.

When all students receive the same items, as implied by the linear model in Equation 25, the main effect due to the sampling of items, $\beta_{\underline{I}}$, affects all examinees' observed scores in the same way. For mastery testing, however, this does not eliminate the item effect as a source of error, because our interest is in the absolute magnitude of an examinee's score, not the magnitude relative to the scores of other examinees. For example, if we happen to select an especially easy set of items from the universe, our estimates of $\mu_{\underline{v}}$ (for the universe of items) will tend to be too high for all examinees; this error is accounted for by $\beta_{\underline{I}}$.

Interpretation of $\theta(\underline{d})$ and $\theta_{\underline{c}}(\underline{d})$

The second block of Table 1 summarizes results for the parameters, $\underline{A}(\underline{d})$, $\underline{A}_{\underline{m}}(\underline{d})$, $\underline{A}_{\underline{c}}(\underline{d})$, and $\underline{L}(\underline{d})$, and the agreement indices, $\theta(\underline{d})$ and $\theta_{\underline{c}}(\underline{d})$, for the domain-referenced agreement function, $\underline{d}$.

The difference in interpretation between $\theta(\underline{d})$ and $\theta_{\underline{c}}(\underline{d})$ parallels the difference between $\theta(\underline{t})$ and $\theta_{\underline{c}}(\underline{t})$. The index $\theta(\underline{d})$ characterizes the dependability of decisions or estimates based on the testing procedures. The index, $\theta_{\underline{c}}(\underline{d})$, indicates the contribution of the testing procedures to the dependability of these decisions or estimates. It is clear from Equation 33 that $\theta(\underline{d})$ may be large even when there is little or no universe score variability in the population of examinees. From Equation 35, however, we see that $\theta_{\underline{c}}(\underline{d})$ is equal to zero when there is no universe score variability in the population [assuming $\sigma^2(\Delta) > 0$] .

Norm-referenced tests compare each examinee's score to the scores of other examinees, and, therefore, require variability if these comparisons are to be dependable. In their now classic paper, Popham and Husek (1969) maintained that "variability is not a necessary condition for a good criterion-referenced test" (p. 3). They argued that since criterion-referenced tests are "used to ascertain an individual's status with respect to some criterion" (p. 2), the meaning of the score is not dependent on comparison with other scores. Popham and Husek conclude, therefore, that indices of dependability that require variability are appropriate for norm-referenced tests but not for criterion-referenced tests.

Although the position adopted by Popham and Husek seems plausible, it leads to a very disturbing conclusion. As Woodson (1974a, p. 64) has pointed out, "items and tests which give no variability...give no information and are

therefore not useful." We are faced, therefore, with the apparent contradiction, or paradox, that tests which provide no information about differences among individual examinees can be good criterion-referenced tests. In two subsequent articles, Millman and Popham (1974) and Woodson (1974b) clarified the two sides of this dispute without resolving the basic issue.

The general framework developed here provides an obvious resolution of this paradox. As we have seen, two types of coefficients can be developed for any agreement function, depending upon whether or not one corrects for chance agreement. Coefficients, such as $\theta(\underline{d})$, that are not corrected for chance provide estimates of the dependability of the decision procedures; and such coefficients may be large even without variability in universe scores. By contrast, coefficients such as $\theta_{\underline{c}}(\underline{d})$, that are corrected for chance provide an estimate of the contribution of the test to the dependability of the decision procedure. Such coefficients will approach zero as the universe score variance approaches zero. Popham and Husek's argument applies to the decision procedure, and coefficients not corrected for chance are appropriate for characterizing the dependability of the decision procedure. Woodson's argument applies to the contribution of the test to the decision procedure, and coefficients corrected for chance are appropriate for characterizing the contribution of the test to the dependability of the decision procedure.

## Domain-Referenced Agreement Without a Cutting Score

The domain-referenced agreement function in Equation 27 is the product of deviations from a constant. The discussion up to this point has focused on mastery testing, and $\lambda$ has been taken as the mastery cutting score.

However, a single domain-referenced test may be used for several different decisions, involving different cutting scores. In such cases, it would be useful to have an index of dependability that does not depend on a particular cutting score. As discussed earlier, $\theta_{\underline{c}}(\underline{d})$ is independent of $\lambda$, and $\theta_{\underline{c}}(\underline{d})$ is appropriate for assessing the contribution made by the test to the dependability of mastery decisions using any cutting score. Furthermore, $\theta_{\underline{c}}(\underline{d})$ is less than or equal to $\theta(\underline{d})$ for all values of $\lambda$, and the two are equal only when $\lambda = \mu$. Therefore, $\theta_{\underline{c}}(\underline{d})$ provides a lower bound for $\theta(\underline{d})$ (see Brennan, 1977b).

Moreover, domain-referenced tests do not necessarily involve any consideration of cutting scores. For example, the score, $\underline{S}_{vI}$, on a domain-referenced test may be interpreted as a descriptive statistic which estimates $\mu_{\underline{v}}$, the examinee's universe score (i.e., percentage of items that could be answered correctly) in the domain (see Millman and Popham, 1974). When using domain-referenced scores as descriptive statistics, we are interested in point estimates of the examinee's universe score, $\mu_{\underline{v}}$. As we have seen, the error (or noise) in such point estimates of universe scores is given by $\Delta_{\underline{v}}$, and $\theta_{\underline{c}}(\underline{d})$ therefore incorporates the appropriate error variance $\sigma^2(\Delta)$. The universe score variance, $\sigma^2(\pi)$ in $\theta(\underline{d})$ provides a measure of the dispersion of universe scores in the population. There is a strong precedent in physical measurement for taking the variability in universe scores as a measure of the magnitude of the signal to be detected. General-purpose instruments for measuring length, for example, are typically evaluated by their ability to detect differences of the order of magnitude of those encountered in some area of practice. Thus, rulers are adequate in carpentry, but verniers are necessary in machine shops.

## $\theta$ and $\theta_c$ for Norm-Referenced Agreement

The agreement function that is implicit in generalizability coefficients (see Cronbach et al, 1972, and Brennan, 1977a) is:

$$g(S_{vI}, S_{wJ}) = (S_{vI} - \mu_I)(S_{wJ} - \mu_J), \tag{37}$$

where $\mu_I$ is the expected value of $S_{vI}$ over the population of persons for the set of items $I$; that is,

$$\mu_I = \mathcal{E}_v S_{vI} = \mu + \beta_I . \tag{38}$$

Similarly,

$$\mu_J = \mathcal{E}_v S_{vJ} = \mu + \beta_J . \tag{39}$$

The parameters, $\beta_I$ and $\beta_J$, are the average values of the item effect for the two samples of items, and they reflect differences in difficulty level from one randomly-selected instance of the testing procedure to another.

Note that the agreement function for norm-referenced tests, given in Equation 37 and the agreement function for domain-referenced tests given in Equation 27 are both products of deviation scores. The difference between the two agreement functions is in the nature of the deviation scores that are used. The norm-referenced agreement function is defined in terms of deviations from the population mean for fixed sets of items. These deviation scores compare the examinee's performance on the set of items to the performance

of the population on the same set of items. The domain-referenced agreement function in Equation 27 is defined in terms of the deviation of the examinee's score from a fixed cutting score.

The Indices $\theta(g)$ and $\theta_c(g)$

Using the norm-referenced agreement function in Equation 37 and the definition of expected agreement in Equation 2, we obtain

$$A(g) = \sum_{v,I,J} [(S_{vI} - \mu_I)(S_{vJ} - \mu_J)] ; \qquad (40)$$

and using Equation 25 to replace $S_{vI}$ and $S_{vJ}$ in Equation 40,

$$A(g) = \sum_{v,I,J} [\pi_v + (\pi\beta,e)_{vI}] \cdot [\pi_v + (\pi\beta,e)_{vJ}] = \sigma^2(\pi). \qquad (41)$$

Similarly, the maximum expected agreement is found by using the norm-referenced agreement function and the definition of maximum expected agreement in Equation 6:

$$A_m(g) = \sum_{v,I} (S_{vI} - \mu_I)^2 = \sigma^2(\pi) + \frac{\sigma^2(\pi\beta,e)}{n}, \qquad (42)$$

where $n$ is the number of items sampled for each instance of the testing procedure. Substituting Equations 41 and 42 in Equation 5, an index of dependability for norm-referenced tests is:

$$\theta(g) = \frac{\sigma^2(\pi)}{\sigma^2(\pi) + \frac{\sigma^2(\pi\beta,e)}{n}} . \qquad (43)$$

Using the norm-referenced agreement function in Equation 37 and the definition of chance agreement in Equation 8, the expected agreement due to chance is:

$$\underline{A}_{\underline{c}}(\underline{g}) = \underset{\underline{v},\underline{w},\underline{I},\underline{J}}{\mathcal{E}} [(\underline{S}_{\underline{vI}} - \mu_{\underline{I}})(\underline{S}_{\underline{wJ}} - \mu_{\underline{J}})]$$

$$= \underset{\underline{v},\underline{w},\underline{I},\underline{J}}{\mathcal{E}} [\pi_{\underline{v}} + (\pi\beta,\underline{e})_{\underline{vI}}] \bullet [\pi_{\underline{w}} + (\pi\beta,\underline{e})_{\underline{wJ}}] .$$

Since all of the effects in this equation are assumed to be sampled independently,

$$\underline{A}_{\underline{c}}(\underline{g}) = 0; \tag{44}$$

and, therefore,

$$\theta_{\underline{c}}(\underline{g}) = \theta(\underline{g}). \tag{45}$$

The correction for chance has no effect on the norm-referenced dependability index, because a correction for chance is built into the norm-referenced agreement function in Equation 37.

Norm-Referenced Loss and $\sigma^2(\delta)$

The loss associated with the norm-referenced agreement function is found by subtracting Equation 41 from Equation 42:

$$\underline{L}(\underline{g}) = \underline{A}_{\underline{m}}(\underline{g}) - \underline{A}(\underline{g}) = \sigma^2(\pi\beta,\underline{e})/\underline{n} . \tag{46}$$

This loss is simply the error variance designated by Cronbach et al (1972) as $\sigma^2(\delta)$, which is also the error variance in classical test theory.

In norm-referenced testing, we are interested in "the relative ordering of individuals with respect to their test performance, for example, whether student A can solve his problems more quickly than student B" (Glaser, 1963, p. 519). Thus, our interest is in "the adequacy of the measuring procedure for making comparative decisions" (Cronbach et al., 1972, p. 95). In this situation, the error for a given person, as defined by Cronbach et al. (1972) is

$$\begin{aligned} \delta_{\underline{v}} &= (S_{\underline{vI}} - \mu_{\underline{I}}) - (\mu_{\underline{v}} - \mu) \\ &= [\mu + \pi_{\underline{v}} + \beta_{\underline{I}} + (\pi\beta,\underline{e})_{\underline{vI}} - \mu - \beta_{\underline{I}}] - [\mu + \pi_{\underline{v}} - \mu] \\ &= (\pi\beta,\underline{e})_{\underline{vI}} . \end{aligned}$$

The variance of $\delta_{\underline{v}}$ over the population of persons and samples of $\underline{n}$ items is

$$\sigma^2(\delta) = \sigma^2(\pi\beta,\underline{e})/\underline{n} = \underline{L}(\underline{g}) . \qquad (47)$$

From Equations 14 and 45

$$\theta(\underline{g}) = \theta_{\underline{c}}(\underline{g}) = \frac{\underline{A}(\underline{g})}{\underline{A}(\underline{g}) + \underline{L}(\underline{g})} ;$$

and substituting Equations 41 and 47

$$\theta(g) = \theta_c(g) = \frac{\sigma^2(\pi)}{\sigma^2(\pi) + \sigma^2(\delta)}, \qquad (48)$$

which is identical to the generalizability coefficient $\varepsilon\rho^2$ given the random effects linear model in Equation 25. [Equation 48 is also equivalent to Cronbach's (1951) coefficient alpha and to KR-20 for dichotomously scored items.]

Interpretation of $\theta(g) = \theta_c(g)$

The third block of Table 1 summarizes results for the parameters, $A(g)$, $A_m(g)$, $A_c(g)$, and $L(g)$, and the agreement indices, $\theta(g)$ and $\theta_c(g)$, for the norm-referenced agreement function, $g$.

Equations 43 and 48 can also be interpreted as an intraclass correlation coefficient, and, as such, they are approximately equal to the expected correlation between random instances of the testing procedure (i.e., independent random samples of $n$ items). Estimation procedures for generalizability coefficients are discussed by Cronbach et al. (1972), and by Brennan (1977a).

From Equations 35 and 43 (or 48), note that $\theta_c(d)$ and $\theta_c(g)$ incorporate the same expected agreement (or signal) but different definitions of error variance (loss or noise). For $\theta_c(d)$ the error variance is $\sigma^2(\Delta)$, and for $\theta_c(g)$ the error variance is $\sigma^2(\delta)$. It follows that

$$\theta_c(d) \leq \theta_c(g)$$

because

$$\sigma^2(\delta) \leq \sigma^2(\Delta).$$

The difference between $\sigma^2(\Delta)$ and $\sigma^2(\delta)$ is simply $\sigma^2(\beta)/n$. Therefore, $\theta_c(d)$ and $\theta_c(g)$ are equal only when $\beta_I$ is a constant for all instances of the testing procedure. The variance component for the main effect for items, $\sigma^2(\beta)$, reflects differences in the mean score (in the population) for different samples of items. If we are interested only in differences among examinee universe scores, as in norm-referenced testing, then any effect which is a constant for all examinees does not contribute to the error variance. However, for domain-referenced testing, we are interested in the absolute magnitude of examinee universe scores, or the magnitude compared to some externally defined cutting score. In this case, fluctuations in mean scores for samples of items do contribute to error variance.

## The Effect of Item Sampling on the Indices of Dependability, $\theta$ and $\theta_c$

### Items Nested within Persons in the D Study

We have examined the implications of using several definitions of agreement for randomly parallel tests. We have assumed that, for each instance of the testing procedure, a random sample of items from some infinite domain is administered to all examinees; i.e., items are crossed with examinees. Following Cronbach et al. (1972), this design is designated p x i. Indices that are appropriate for other designs can be derived using the approach discussed above. A particularly interesting and useful set of indices is obtained by assuming that an independent random sample of items is selected for each examinee. Following Cronbach et al. (1972), this design is designated i:p, where the colon means "nested within."

In this section it will be convenient to make use of the distinction between a G study and a D study--a distinction originally drawn by Rajaratnam (1960) and subsequently discussed extensively by Cronbach, et al. (1972). The purpose of a G study, or generalizability study, is to examine the dependability of some measurement procedure. The purpose of a D study, or decision study, is to provide the data for making substantive decisions. "For example, the published estimates of reliability for a college aptitude test are based on a G study. College personnel officers employ these estimates to judge the accuracy of data they collect on their applicants (D study)" (Cronbach, et al., 1972, p. 16). The principal results of a G study are estimates of variance components, which can then be used in a variety of D studies. The G study and the D study may use the same design or different designs. Generally, G studies are most useful when they employ crossed designs and large sample sizes to provide stable estimates of as many variance components as possible.

In previous sections of this paper, we have implicitly assumed that both the G study and the D study used the crossed design, $\underline{p} \times \underline{i}$. We will continue to assume that variance components have been estimated from the crossed design. However, in this section we will assume that the D study employs the $\underline{i}:\underline{p}$ design. For example, in computer-assisted testing it is frequently desirable (or even necessary for security reasons) that each examinee receive a different set of items; i.e., the D study uses an $\underline{i}:\underline{p}$ design. However, even in such cases it is desirable that the variance components be estimates from the crossed design, $\underline{p} \times \underline{i}$.

If in the D study each examinee gets a different set of items, the item effect will not be the same for all examinees. Under these circumstances, the linear model for scores on a particular instance of the testing procedure is:

$$\underline{S}_{\underline{vI}} = \mu + \pi_{\underline{v}} + \beta_{\underline{vI}} + (\pi\beta,\underline{e})_{\underline{vI}} ; \qquad (49)$$

and the item effect is now confounded with the residual, $(\pi\beta,\underline{e})_{\underline{vI}}$. It is particularly important to note that, for Equation 49,

$$\mathop{\mathcal{E}}_{\underline{v}} \underline{S}_{\underline{vI}} = \mu ,$$

where $\mu$ is the grand mean in the population of persons and the universe of items. The population mean of the observed scores, $\underline{S}_{\underline{vI}}$, does not equal $\mu_{\underline{I}}$, which is the expected value over the population for a particular set of items, $\underline{I}$. When items are nested within persons, taking the expected value of the observed scores over the infinite population of examinees implies taking the expected value over an infinite universe of items.

## Implications for Norm-Referenced and Domain-Referenced Indices of Dependability

Using the linear model in Equation 49 and the norm-referenced agreement function in Equation 37, it can be shown that:

$$\underline{A}(\underline{g}') = \sigma^2(\pi) , \qquad (50)$$

$$\underline{A}_{\underline{m}}(\underline{g}') = \sigma^2(\pi) + \frac{\sigma^2(\beta)}{\underline{n}} + \frac{\sigma^2(\pi\beta,\underline{e})}{\underline{n}} , \qquad (51)$$

$$\text{and} \quad \underline{A}_{\underline{c}}(\underline{g}') = 0 , \qquad (52)$$

where the prime following $\underline{g}$ differentiates quantities associated with the nested design, $\underline{i}:\underline{p}$, from quantities associated with the crossed design, $\underline{p} \times \underline{i}$.

Substituting these results in Equations 5 and 12, we obtain

$$\theta(\underline{g}') = \theta_{\underline{c}}(\underline{g}') = \frac{\sigma^2(\pi)}{\sigma^2(\pi) + \dfrac{\sigma^2(\beta) + \sigma^2(\pi\beta,\underline{e})}{\underline{n}}} \tag{53}$$

$$= \frac{\sigma^2(\pi)}{\sigma^2(\pi) + \sigma^2(\Delta)} \, . \tag{54}$$

Note that both $\theta(\underline{g}')$ and $\theta_{\underline{c}}(\underline{g}')$ are identical to $\theta_{\underline{c}}(\underline{d})$, the domain-referenced dependability index, corrected for chance, in Equation 35. The only difference between $\theta(\underline{g}')$ and the usual dependability index for norm-referenced tests, $\theta(\underline{g})$, is that $\theta(\underline{g}')$ has an additional term, $\sigma^2(\beta)/\underline{n}$, in the denominator.

For norm-referenced tests, when the same items are administered to all examinees, the item effect, $\beta_{\underline{I}}$, is a constant for all examinees, and $\sigma^2(\beta)/\underline{n}$ does <u>not</u> enter the error variance. If items are nested within examinees, however, $\beta_{\underline{vI}}$, will generally be different for each examinee, and $\sigma^2(\beta)/\underline{n}$ is part of the error variance.

For the domain-referenced agreement function, the agreement indices developed from the nested model are identical to those developed from the crossed model:

$$\theta(\underline{d}') = \theta(\underline{d})$$

and

$$\theta_{\underline{c}}(\underline{d}') = \theta_{\underline{c}}(\underline{d}).$$

The dependability of a domain-referenced testing procedure is not affected by whether the D study uses the crossed design, $p \times i$, or the nested design $i:p$. The aim of domain-referenced testing is to provide point estimates of examinee universe scores, rather than to make comparisons among examinees. The dependability of each examinee's score is determined by the number of items administered to that examinee, not by how many items or which items are administered to other examinees.

Standardization of the items used in any instance of the testing procedure improves the dependability of norm-referenced interpretations but does not improve the dependability of domain-referenced interpretations. Furthermore, the use of different samples of items for different examinees will tend to improve estimates of group means. If, therefore, domain-referenced tests are to be used for program evaluation, the selection of independent samples of items for different examinees provides more dependable estimates of group means without any loss in the dependability of estimates of examinees' universe scores.

## Summary and Conclusions

Table 1 provides an overview of the major results derived and discussed in this paper. Two indices of dependability, $\theta$ and $\theta_c$, are discussed for three different agreement functions: the threshold agreement function, $t$, the domain-referenced agreement function, $d$, and the norm-referenced agreement function, $g$. This paper emphasizes considerations relevant to the first two agreement functions, because the indices of dependability associated with them are indices that have been proposed for domain-referenced and mastery tests. The norm-referenced agreement function, $g$, is considered primarily for purposes of comparing it with the other two agreement functions. The main purposes of this generalized treatment of indices of dependability are to provide an internally consistent framework for deriving indices of dependability for domain-referenced tests, and to examine the implications of choosing a particular index.

### Choosing an Index of Dependability

Our discussion of these issues has not dictated which index an evaluator should choose in a particular context, but our discussion has indicated that two main issues are involved in such a choice: (a) the nature of agreement functions (or, alternatively, loss functions), and (b) the use of an index corrected for chance or not corrected for chance.

With respect to the first issue, two types of agreement functions have been considered for mastery tests: the threshold agreement function (Equation 16 or 17) and the domain-referenced agreement function (Equation 27). The threshold agreement function is appropriate whenever the only distinction that can be made usefully is a qualitative distinction between masters and

non-masters. If, however, different degrees of mastery and non-mastery exist to an appreciable extent, the threshold agreement function is not appropriate because it ignores such differences.

In most educational contexts, differences between masters and non-masters are not purely qualitative. Rather, the attribute that is measured is conceptualized as an ordinal or interval scale, and the examinees may possess the attribute to varying degrees even though a single cutting score is used to define mastery. In this context it is important that examinees who are far above or below the cutting score be classified correctly. The misclassification of such examinees is likely to cause serious losses. The misclassification of examinees whose level of ability is close to the cutting score will involve much less serious losses. Current techniques for setting the cutting score are not very precise, and the choice of a cutting score, is to some extent, arbitrary. It is, therefore, relatively less important that the testing procedure correctly classify examinees whose level of skill is close to the specified cutting score.

The domain-referenced agreement function, $\underline{d}$, in Equation 27 reflects these considerations. It assigns a positive value to the agreement whenever both instances of the testing procedure place the examinee in the same category, and it assigns a negative value to the agreement when the two instances place an examinee in different categories. Furthermore, the magnitude of the agreement is determined by the distance of the observed scores from the cutting score on the two instances of the procedure.

The second issue in choosing an index of dependability is whether to use the index $\theta$, which is not corrected for chance agreement, or the index $\theta_{\underline{c}}$,

which is corrected for chance. There is no reason to prefer one index over the other in all contexts. The two indices provide different information, and, therefore, should be interpreted differently. For judgements about the dependability of a decision procedure, as applied to a particular population, indices that are not corrected for chance are more appropriate. For judgements about the contribution of tests to the dependability of the decision procedure, indices that are corrected for chance are more appropriate. Subkoviak (Note 3) makes similar statements in his response to Huynh's (Note 5) criticism of coefficient kappa [$\theta_c(t)$ given by Equation 23].

It is also useful to note that whether one chooses $\theta$ or $\theta_c$, the expected loss or error variance remains unchanged. That is, the choice between $\theta$ and $\theta_c$ usually affects the strength of the signal in a testing procedure, but never the strength of the noise (see Brennan and Kane, 1977b). In effect, when one chooses $\theta_c$, the strength of the signal is reduced by an amount attributable to chance, and it is this reduction of signal strength that causes $\theta_c$ to be less than $\theta$, usually. As noted previously, for the norm-referenced agreement function, $g$, $\theta$ always equals $\theta_c$ because chance agreement is zero. Indeed, this is probably one reason why the distinction between indices such as $\theta$ and $\theta_c$ has been ignored in much of the literature on testing and psychometrics.

## Prior Information

For the domain-referenced agreement function, $d$, $\theta(d)$ equals $\theta_c(d)$ when $(\mu - \lambda)^2$ equals zero, i.e., when the mean, $\mu$, equals the cutting score, $\lambda$. In such cases, prior information about $\mu$ is of no use in classifying examinees as masters, or non-masters; and the dependability of decisions depends entirely upon the dependability of the test being used. If $(\mu - \lambda)^2$ is very large,

decisions made about a student's mastery or non-mastery status, solely on the basis of prior information about $\mu$, may be highly dependable. If, however, $(\mu - \lambda)^2$ is non-zero but not very large compared to the expected loss, $\sigma^2(\Delta)$, it is likely that the dependability of decisions could be improved by using Bayesian methods.

Bayesian procedures (Hambleton and Novick, 1973; Swaminathan, Hambleton, and Algina, 1975) take advantage of prior information about the population by using this information and the student's observed score to estimate the student's universe score. The optimum weighting of prior information and test scores depends on the prior distribution of universe scores in the population, the dependability of the testing procedure, and the agreement function (or equivalently, the loss function) that is chosen. Although the published applications of Bayesian methods have used a threshold loss, these methods are, in principle, equally applicable for the domain-referenced loss, $\sigma^2(\Delta)$.

Assumptions about Parallel Tests

Throughout this paper, we have assumed that two tests are parallel if they involve random samples of the same number of items from the same universe, or domain, of items. That is, we have made the assumption of randomly-parallel tests, rather than the stronger assumption of classically parallel tests. Cronbach et al. (1972) have shown that either assumption can be used as a basis for defining the generalizability coefficient for the persons crossed with items design; and we have shown that this generalizability coefficient is identical to $\theta(g) = \theta_c(g)$ for norm-referenced tests. Also, either assumption, in conjunction with the threshold agreement function, can

be used to derive the indices $\theta(\underline{t})$ and $\theta_{\underline{c}}(\underline{t})$. It is interesting to note, however, that the Huynh (1976) and Subkoviak (1976) procedures for estimating $\theta(t)$ and $\theta_{\underline{c}}(t)$ necessitate the assumption of classically parallel tests.

We have argued that the assumption of classically parallel tests is generally inappropriate for a domain-referenced test because, for a domain-referenced test, our interest is focused on an examinee's universe score without regard to the scores of other examinees. However, if all items in the universe are equally difficult for the population of persons, then the item effect, $\beta_{\underline{i}}$, in Equation 24 is a constant for all items, and $\sigma^2(\Delta)$ equals $\sigma^2(\delta)$. That is, the expected loss for the domain-referenced agreement function equals the expected loss for the norm-referenced agreement function. In this case, the index $\theta(\underline{d})$ in Equation 33 is identical to Livingston's (1972a, 1972b, 1972c, 1973) coefficient.

The differences between $\theta(\underline{d})$ and Livingston's coefficient are, therefore, a direct result of the differences between the assumptions of randomly parallel tests and classically parallel tests, respectively. It is important to note, however, that neither index is corrected for chance. They both reflect the dependability of a decision procedure, not the contribution of tests to the dependability of a decision procedure. Also, for both coefficients changes in the cutting score, $\lambda$, affect the coefficients' magnitudes through the signal strength, not through the noise or error variance.

## Concluding Comments

Throughout this paper we have concentrated upon indices of dependability for domain-referenced tests, and factors that influence the use and interpretation of such indices. We have particularly emphasized the indices $\theta(\underline{d})$ and $\theta_{\underline{c}}(\underline{d})$ because they have broad applicability in domain-referenced testing, they are easily compared with the usual norm-referenced indices of dependability, and they can be developed using principles from generalizability theory--a broadly applicable psychometric model. Using principles from generalizability theory, it is relatively straightforward to define $\theta(\underline{d})$ and $\theta_{\underline{c}}(\underline{d})$ for ANOVA designs other than the simple persons-crossed-with-items design. (See, for example, our treatment of the items nested within persons design.) The extension of $\theta(\underline{t})$ and $\theta_{\underline{c}}(\underline{t})$ to other designs is not so straightforward.

However, no matter which index of dependability an evaluator chooses, it is important that the evaluator recognize the underlying assumptions and interpret results in a meaningful manner. In this regard, it is often the case that the magnitude of an index of dependability, alone, provides an insufficient basis for decision-making. It is almost always best to provide, also, the quantities that enter the index ($\underline{A}$, $\underline{A}_{\underline{m}}$, $\underline{A}_{\underline{c}}$, and $\underline{L}$ in Table 1), as well as the estimated variance components (see APA, 1974).

## Reference Notes

1. Marshall, J. L., & Haertel, E. H. The mean split-half coefficient of agreement: A single administration index of reliability for mastery tests. Unpublished manuscript, 1976. (Available from Department of Educational Psychology, 1025 West Johnson Street, Madison, Wisconsin 53706).

2. Berger, R. J. A measure of reliability for criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Minneapolis, March 1970.

3. Subkoviak, M. J. Further comments on reliability for mastery tests (Laboratory of Experimental Design, Occasional Paper No. 17). Unpublished manuscript, University of Wisconsin, 1977.

4. Subkoviak, M. J., & Albrecht, B. A. Empirical investigation of procedures for estimating reliability for mastery tests. Unpublished manuscript, 1977. (Available from Department of Educational Psychology, 1025 West Johnson Street, Madison, Wisconsin 53706).

5. Huynh, H. Reliability of criterion-referenced tests: Comments on a paper by Subkoviak. Manuscript submitted for publication, 1977.

References

American Psychological Association. Standards for educational & psychological tests (rev. ed.). Washington, D. C.: American Psychological Association, 1974.

Brennan, R. L. Generalizability analyses: Principles and procedures. ACT Technical Bulletin No. 26. Iowa City: The American College Testing Program, September 1977. (a)

Brennan, R. L. KR-21 and lower limits of an index of dependability for mastery tests. ACT Technical Bulletin No. 27. Iowa City: The American College Testing Program, December 1977. (b)

Brennan, R. L., & Kane, M. T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, 14, 277-289. (a)

Brennan, R. L., & Kane, M. T. Signal/noise ratios for domain-referenced tests. Psychometrika, 1977, 42, . (b)

Carver, R. P. Special problems in measuring change with psychometric devices. In Evaluative research: Strategies and methods. Pittsburgh: American Institutes for Research, 1970.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 292-334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 1963, 18, 519-521.

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Hively, W. Introduction to domain-referenced testing. In W. Hively (Ed.), Domain-referenced testing. Englewood Cliffs, N.J.: Educational Technology Publications, 1974.

Huynh, H. On consistency of decisions in criterion-referenced testing. Journal of Educational Measurement, 1976, 13, 265-275.

Ivens, S. H. An investigation of item analysis, reliability, and validity in relation to criterion-referenced tests. Unpublished doctoral dissertation, Florida State University, August 1970.

Kriewall, T. E. Application of information theory and acceptance sampling principles to the management of mathematics instruction. (Doctoral dissertation, University of Wisconsin) Ann Arbor, Michigan: University Microfilms, 1969, No. 69-22,417.

Livingston, S. A. A criterion-referenced application of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26. (a)

Livingston, S. A. A reply to Harris' "An interpretation of Livingston's reliability coefficient for criterion-referenced tests." Journal of Educational Measurement, 1972, 9, 31. (b)

Livingston, S. A. Reply to Shavelson, Block, and Ravitch's "Criterion-referenced testing: Comments on reliability." Journal of Educational Measurement, 1972, 9, 139. (c)

Livingston, S. A. A note on the interpretation of the criterion-referenced reliability coefficient. Journal of Educational Measurement, 1973, 4, 311.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education. Berkeley, Calif.: McCutchan, 1974.

Millman, J., & Popham, W. J. The issue of item and test variance for criterion-referenced tests: A clarification. Journal of Educational Measurement, 1974, 11, 137-138.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Rajaratnam, N. Reliability formulas for independent decision data when reliability data are matched. Psychometrika, 1960, 25, 261-271.

Scott, W. A. Reliability of content analysis: The case of nominal scale coding. Public Opinion Quarterly, 1955, 19, 321-325.

Subkoviak, M. J. Estimating reliability from a single administration of a mastery test. Journal of Educational Measurement, 1976, 13, 253-264.

Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-267.

Swaminathan, H., Hambleton, R. K., & Algina, J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. Journal of Educational Measurement, 1975, 12, 87-98.

Woodson, M. I. The issue of item and test variance for criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 63-64. (a)

Woodson, M. I. The issue of item and test variance for criterion-referenced tests: A reply. Journal of Educational Measurement, 1974, 11, 139-140. (b)

## Footnote

This research was partially supported by ONR Contract No. N00123-77-C-0739 between the American College Testing Program and the Navy Personnel Research and Development Center.

Table 1

Coefficients for Different Agreement Functions

| Agreement Function | Parameters | Agreement Coefficients |
|---|---|---|
| Threshold: | $\underline{A}(\underline{t}) = \Sigma \underline{p}_{\underline{ii}}$ | $\theta(\underline{t}) = \Sigma \underline{p}_{\underline{ii}}$ |
| $\underline{t}(\underline{S}_{\underline{vk}}, \underline{S}_{\underline{wl}})$ | $\underline{A}_{\underline{m}}(\underline{t}) = 1$ | |
| $= \begin{cases} 1 \text{ if } \underline{S}_{\underline{vk}} = \underline{S}_{\underline{wl}} \\ 0 \text{ if } \underline{S}_{\underline{vk}} \neq \underline{S}_{\underline{wl}} \end{cases}$ | $\underline{A}_{\underline{c}}(\underline{t}) = \Sigma \underline{p}_{\underline{i}}^2$ | $\theta_{\underline{c}}(\underline{t}) = \dfrac{\Sigma \underline{p}_{\underline{ii}} - \Sigma \underline{p}_{\underline{i}}^2}{1 - \Sigma \underline{p}_{\underline{i}}^2}$ |
| | $\underline{L}(\underline{t}) = \Sigma \underline{p}_{\underline{ij}} \ (\underline{i} \neq \underline{j})$ | |
| Domain-Referenced: | $\underline{A}(\underline{d}) = (\mu - \lambda)^2 + \sigma^2(\pi)$ | |
| $\underline{d}(\underline{S}_{\underline{vI}}, \underline{S}_{\underline{wJ}})$ | $\underline{A}_{\underline{m}}(\underline{d}) = (\mu - \lambda)^2 + \sigma^2(\pi) + \sigma^2(\Delta)$ | $\theta(\underline{d}) = \dfrac{(\mu - \lambda)^2 + \sigma^2(\pi)}{(\mu - \lambda)^2 + \sigma^2(\pi) + \sigma^2(\Delta)}$ |
| $= (\underline{S}_{\underline{vI}} - \lambda)(\underline{S}_{\underline{wJ}} - \lambda)$ | $\underline{A}_{\underline{c}}(\underline{d}) = (\mu - \lambda)^2$ | |
| | $\underline{L}(\underline{d}) = \sigma^2(\Delta)$ | $\theta_{\underline{c}}(\underline{d}) = \dfrac{\sigma^2(\pi)}{\sigma^2(\pi) + \sigma^2(\Delta)}$ |
| Norm-Referenced: | $\underline{A}(\underline{g}) = \sigma^2(\pi)$ | |
| $\underline{g}(\underline{S}_{\underline{vI}}, \underline{S}_{\underline{wJ}})$ | $\underline{A}_{\underline{m}}(\underline{g}) = \sigma^2(\pi) + \sigma^2(\delta)$ | $\theta(\underline{g}) = \theta_{\underline{c}}(\underline{g}) = \dfrac{\sigma^2(\pi)}{\sigma^2(\pi) + \sigma^2(\delta)}$ |
| $= (\underline{S}_{\underline{vI}} - \mu_{\underline{I}})(\underline{S}_{\underline{wJ}} - \mu_{\underline{J}})$ | $\underline{A}_{\underline{c}}(\underline{g}) = 0$ | |
| | $\underline{L}(\underline{g}) = \sigma^2(\delta)$ | |